The Backwards Induction Paradox

Ryan Doody

The Prisoners' Dilemma

In general, a Prisoners' Dilemma (PD) is a game with the following structure:

		Player B	
		c	d
Player A	C	3, 3	0, 4
	D	4, 0	1, 1

Player 1 has the following preferences:

$$(D \wedge c) \succ (C \wedge c) \succ (D \wedge d) \succ (C \wedge d)$$
$$4 > 3 > 1 > 0$$

And Player 2 has the following preferences:

$$(C \wedge d) \succ (C \wedge c) \succ (D \wedge d) \succ (D \wedge c)$$
$$4 > 3 > 1 > 0$$

Both players have a *dominant strategy*: a strategy that is guaranteed to result in a better payoff no matter what the other player does. (But the result of both players playing their dominant strategy is an outcome that is *Pareto-dominated* by some other.)

Player 1: D dominates C.

Player 2: d dominates c.

Outcome $(C \wedge c)$ pareto-dominates outcome $(D \wedge d)$.

Repeated Prisoners' Dilemmas

In a one-shot PD, both players should *defect*. But what if the players know that they will play the game again and again? As we saw, there are strategies—e.g., *Tit-for-Tat*—that do much better than *Always Defect*.

Of course, both players would do better by playing *cooperate* in the one-shot PD, too. But, in the repeated game:

> By playing such a strategy, a player signals his willingness to cooperate provided his partner will do the same, and if the signal is read correctly, it will be in the partner's interest to cooperate—in the case of [*Tit-for-Tat*], in every round except the last.

Suppose that the two players will by n PD against each other.

- If both play *Tit-for-Tat*, they will each win $3n$.
- If both play *Always Cheat*, they will each win n.

If one plays *Tit-for-Tat* while the other plays *Always Cheat*, the former wins $0 + (n-1) = n-1$ and the latter wins $4 + (n-1) = n+3$.

But, if both players are rational (and their rationality is a matter of common belief), they are in a position to run the following **Backwards Induction** argument:

(1a) My partner, being rational, will *defect* in the nth round.

> ...since defecting at that stage will not have any undesirable effects in further rounds—there are none.
>
> ...since *defect* will dominate *cooperate*, just as in the one-shot PD.

(1b) My partner will also expect me to *defect* in the nth round.

> ...since they believe that I am rational.

(2a) My partner will *defect* in round $n-1$.

A proposition p is a matter of **common belief** among a group just in case:

(i) every member believes that p,

(ii) every member believes that every member believes that p,

(iii) every member believes that every member believes that every member believes that p,

⋮

...and so on.

. . . since they expect me to defect in round n, and so there will be no undesirable effects in further rounds—the nth result is already fixed.

. . . since defecting will dominate cooperation.

(2b) My partner will also expect me to do likewise in round $n-1$.

⋮

(n) My partner, being rational, will *defect* in the 1st round—and, thus, so will I.

$\therefore$ *Always Defect* is the uniquely rational strategy for each player.

. . . and so on and so forth . . . until we get back to the very first round. So, both players should defect in every round!

But is that right? Is it *really* irrational to play *Tit-for-Tat* in a finite repeated game?

Resolving the Paradox

The **Backwards Induction** involves assuming that, at each subsequent round, your partner will believe that you are rational—irrespective of how you have acted in the interim.

But any act of cooperation (at any stage) will cause the common belief in rationality to breakdown.

And that means that the players are not in a position to run the backwards induction argument—because one of the beliefs required to to run the induction (that the common belief in rationality would survive even if someone were to *cooperate*) isn't one they can rationally hold.

And so the argument that *Always Defect* is the uniquely rational strategy for each player is unsound.

We've assumed common belief of rationality.

But, just because my partner *now* believes that I am rational (and that I believe that they are, and that they believe that I believe that they are, and so on), it doesn't follow that, in round $n-1$, they will *still* believe that I am rational.

(Furthermore, just because my partner *now* believes that I *now* believe them to be rational, it doesn't follow that, in round $n-2$, they will believe that, in round $n-1$, I will *still* believe that they are rational. And so on.)

Rational Cooperation?

Intuitively, it can be rational to *cooperate* in the 1st round. Is that correct?

Yes—but what it's rational to do depends on what A believes B will come to believe about A if A cooperates in the 1st round.

Are there beliefs a player might hold—that are consistent with the common belief in rationality—that would make it rational to *cooperate* in the 1st round?

Note: It's not rational to play *Tit-for-Tat*. It recommends cooperating in the *last* round (so long as the other player cooperated in round $n-1$), but it's *not* rational to cooperate in the last round.

A's rational strategy:	**What A believes B will believe *if* A cooperates in 1st round:**
1. *Tit-for-Tat minus 2*	"A is irrational, and is playing *Tit-for-Tat* until the end. So, I should play *Delayed Tit-for-Tat minus 1*."
2. *Tit-for-Tat minus 4*	"A is rational and believes I am too. But A believes that I believe that A is irrational—and that, thus, I will play *Delayed Tit-for-Tat minus 1*. Hence, A will play *Tit-for-Tat minus 2*. And so I should play *Delayed Tit-for-Tat minus 3*."
3. *Tit-for-Tat minus 6*	"A is rational and believes I am too, and he believes that I believe he is rational. But he believes that I believe that he believes that I believe the he is irrational—and that, thus, I will play *Delayed Tit-for-Tat minus 3*. Hence, A will play *Tit-for-Tat minus 4*. And so I should play *Delayed Tit-for-Tat minus 5*."
⋮	⋮

Conclusion: There's no uniquely rational strategy to adopt.

. . . It depends on what the players believe about each other, and that's underdetermined.